数据和技术运用视角下我国数字人文“数字型”研究现状及启示*

■ 石静　李阳
南京大学信息管理学院　南京 210023

摘　要：[目的/意义]近年来国内数字人文领域成果涌现，并逐渐从认知探索阶段走向数字化探索阶段。从微观的“数据资源”“数字技术”视角出发，通过量化分析来刻画“数字”与“人文”碰撞的真实全貌，以期对数字人文“数字型”研究有一个直观和实际的认知和思考。[方法/过程]利用文献计量、质性分析、多维尺度分析等方法，从研究阵地分布、学科交叉特征、“数字”运用情况、主要研究类型等维度，对国内数字人文相关研究进行深度分析，并提出相关思考。[结果/结论]从研究阵地看，数字人文研究分布均衡、百家争鸣，在多学科具有广阔空间；从数据基础看，少数几个大型专题数据库和研究者自整理数据是主要数据来源；从研究类型看，资源建设型、工具驱动型和场景融合型三类研究各有侧重，场景融合型研究在“数字”和“人文”中较好地实现平衡。未来，需要积极搭建相关数据基础设施，培育学者数字学术能力，推动“人文”角色转变，以支持数字人文研究范式变革。

关键词：数字人文　学科交叉　数据资源　新技术　人文价值
分类号：G251
DOI：10.13266/j.issn.0252－3116.2021.21.021

数字人文（Digital Humanities，DH）源于人文计算（Humanities Computing）[1]，是计算机或数字技术与人文学科交叉的学术活动领域，也可以被定义为进行人文学术研究的新方式。它的显著特征是培养人文和数字之间的双向关系：既使用“数字”来解答“人文”问题，又使用“数字”来提出新的“人文”问题，而这两者往往是同时进行的。

数字人文在我国已引起广泛关注。早在 2003 年，《信息化进程的前沿探索——数字化与 21 世纪人文精神研讨会综述》一文，引用我国著名哲学家陈志良教授的论点，首次使用了数字人文这一名词[2]。2016 年，北京大学举办首届“数字人文论坛”，之后数字人文浪潮在国内爆发，一时间涌现了大量研究，主要集中在数字人文的基本理论、技术驱动下的人文学术实践转向、新合作模式引发的人文学术文化变革以及面向数字人文研究的基础设施建设等方面[3]。2018 年，“大数据视域下数字人文研究”入选年度中国十大学术热点，历史学、文学、语言学、图书情报学、艺术学等多个学科的学者都加大了数字人文研究力度。但随着研究的推进，诸如研究定位不清、理论内核缺失、技术追崇、知识生产“快捷化”等问题日渐凸显，开始有学者提出质疑：“数字”到底为“人文”带来了什么？“数字”在人文学科中的角色与定位如何[4]？学理和实践领域关于这些问题的争论莫衷一是。必须承认，“数字”与“人文”的碰撞为人文学科的繁荣发展带来了新的思考空间和分析问题的方式，过去人文研究一些不能研究的、或者难以研究的，往往都能在数字化支持下得到解决或给与启示。

近几年来，在开放数据、大数据、智能化等思维和环境的影响下，借助数字技术、数据资源和数字基础设施等探讨传统或新颖人文问题的研究大量增加，显现出新的趋势。首先，我国数字人文研究正在从认知探索走向数字化探索，即逐渐从以基础理论、综述评论、发展导向等为主体的思考型研究，逐渐转向为以实际问题为中心、以数字数据资源为利器的“数字型”研究。其次，在这种“数字型”研究中，也逐渐从早期“只

* 本文系国家社会科学基金重大项目“人文社科专题数据库建设规范化管理研究”（项目编号：18ZDA326）研究成果之一。
作者简介：石静（ORCID：0000－0003－1863－1559），博士研究生，E-mail：shijing11@smail.nju.edu.cn；李阳（ORCID：0000－0002－4479－969X），副教授，博士。
收稿日期：2021－03－28　修回日期：2021－07－26　本文起止页码：141－150　本文责任编辑：杜杏叶

重数据”的工程思维，转为“数据为人文服务”的研究思维。基于上述特色与转型趋势，有必要对我国数字人文相关“数字型”研究做深度梳理与分析。为此，不同于大多“概论式”综述，本文将视角聚焦到数字人文的基础要素“数据”与“技术”，将考察范围限定为相关“真实性”学术研究，以期揭示“数字”与“人文”碰撞的真实全貌，帮助“新入者”对此类研究有一个直观和实在的了解，为国内数字人文新趋势的发展提供借鉴。

1 数据来源与研究方法

1.1 数据来源

基于上述研究问题，本文将数据范围限定为国内真正运用数字技术、数字资源去探讨人文问题的科学文献。符合纳入标准的文献需要同时具备以下特征：①使用人文数据资源或数字工具开展研究；②研究目的是解答人文领域的学术问题。为保证查全率和查准率，本文采取主题词检索和人工筛选结合的方式进行文献选取。首先，在中国知网数据库中，以检索式“SU =（数字人文）”进行检索；文章类型包括全部期刊、会议文章；截至 2020 年 12 月 31 日，共获得文献 928 篇。之后，浏览文献标题和摘要，人工剔除不符合三条纳入标准中任意一条的文献，最终得到文献 147 篇，历年发文数量如图 1 相关研究的文献数量年度分布情况见图 1。可以看出，2012 年左右数字人文“数字型”研究开始出现；2017 年之后，文章数量呈快速上升趋势。最早的一项研究属于当时国外发展较成熟的计算语言学方向：黄敏聪学者利用 55 组基本汉字在书籍中的使用频率，计算分析了汉字简繁体的演变过程[5]。

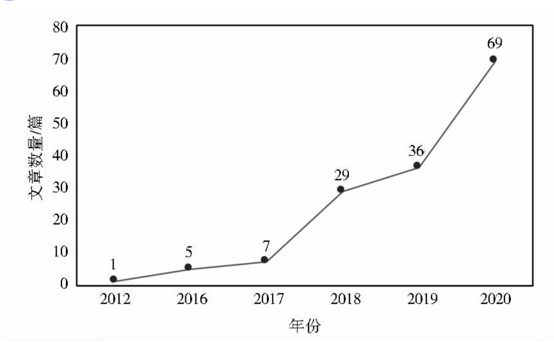


图 1 相关研究的文献数量年度分布情况

1.2 研究方法

本文结合文献计量、质性分析与多维尺度分析的方法开展研究。首先，基于文献的题录信息，对相关研究进行基本描绘分析，包括研究机构分布、学科分布情况和跨学科特征等。之后，基于文献的内容信息，利用质性分析方法对文章的研究内容进行编码标注，在此基础上对数据及方法使用情况进行分类讨论。最后，利用高频关键词开展共词分析，总结主要研究类型。

2 基本描绘分析

2.1 机构与学科分布

统计文章作者所属二级单位，见表 1。武汉大学信息管理学院/信息资源研究中心是首要研究阵地，2011 年便建立了数字人文研究中心，自 2018 年起相关研究逐渐增多，侧重图像资源建设[6]、语义表示方法[7]、知识融合技术[8]等方向，诸如敦煌壁画和文化遗产项目都产生了较大影响。上海图书馆排名第二，其自 2016 年起便有关于家谱数字人文服务的相关研究[8]，在此之后，又构建了董其昌数据库[9]、人名规范库、印章知识库等多个专题知识库，在学术研究中利用率较高。整体来看，机构分布具有以下特点：①仍是以高校的各级学科科研单位为主，占比70.78%；②研究阵地相对分散且均衡，大有“百家争鸣”之势，其中，信息管理学院、经济管理学院、信息科学学院等二级单位占比较高；③数字人文为图书馆的实践工作者参与学术研究提供了新兴路径，相关研究数量占比达到 23.38%，充分反映了数字人文研究中理论与实践紧密结合的特征。

表 1 研究机构分布情况

作者单位	文章数量/篇	作者单位	文章数量/篇
武汉大学信息管理学院/信息资源研究中心	12	上海师范大学图书馆	4
上海图书馆	10	山东大学历史文化学院	4
南京农业大学信息科学技术学院	7	北京大学信息管理系	4
上海交通大学图书馆	6	上海大学图书情报档案系	3
南京大学信息管理学院	6	南通大学经济与管理学院	3
华东师范大学图书馆	6	南京师范大学文学院	3
南京农业大学中华农业文明研究院	5	南京农业大学马克思主义学院	3
华东师范大学经济与管理学部	5	南京大学历史学院	3
中国人民大学信息资源管理学院	4	吉林大学管理学院	3

进一步地，从相关文章的刊载期刊关联到一级学科，探讨相关研究的学科分布情况，结果如图 2。

总体而言，学科分布多元化，共涉及 23 个一级学科，覆盖了人文与社会科学、自然科学、工程与技术科学、农业科学等全部门类，跨学科特征明显。其中，图书馆、情报与文献学相关研究占比最高，其主要贡献在

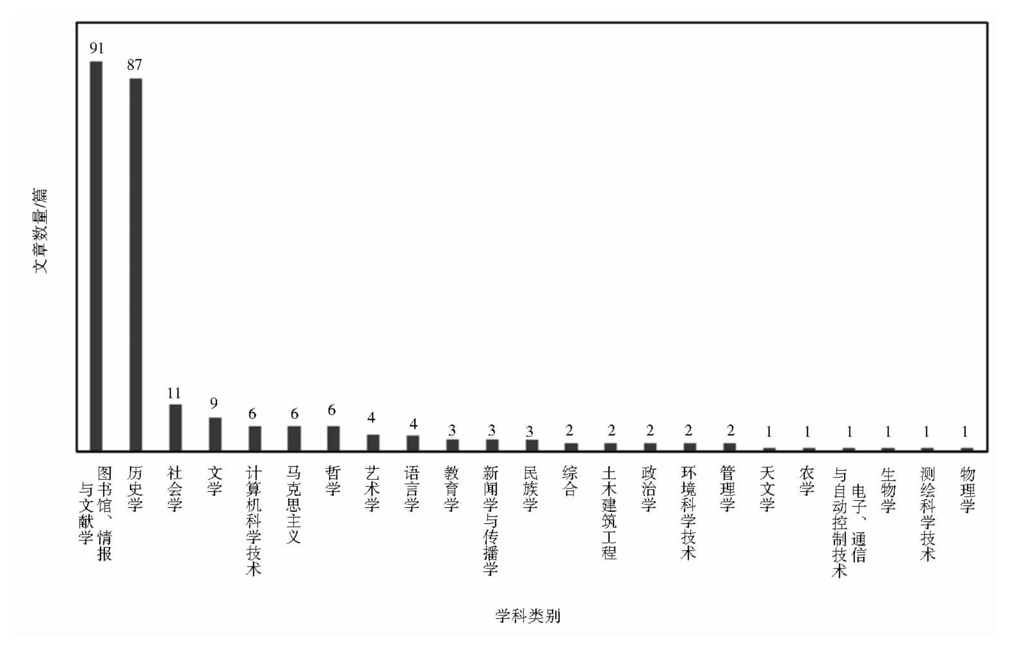


图 2　相关研究的学科分布

于技术、方法和工具的提供,专注于信息资源管理全流程的技术方法,如信息获取中的信息检索[12]、信息描述中的语义表示[7]、信息组织中的元数据[13]、信息挖掘中的数据分析、信息利用中的可视化[14－16]等。之后依次是历史学、社会学和文学,在 2017 年后相关研究增多。虽然数字人文的概念由来已久,但早期的人文史料数字化多在实践中开展;2018 年后,随着数字人文的爆发,人文学者才越来越多地参与研究。从数量来看,历史学研究远远多于社会学和文学,“量化历史”“数字历史”等已经成为重要研究分支。在非人文社科的门类中,计算机科学技术相关研究数量占比最高,这部分学者的主要贡献在于数据资源建设和使用过程中技术难点的攻克。与图情档学科不同,计算机学科所关注的技术问题更加细化、针对性也更强,例如不确定数据的处理[10]、多特征知识融合[11]。

2.2　跨学科性

数字人文自诞生之日起,就携带着跨学科的基因。本节基于作者所属二级单位的一级学科,对每篇研究涉及的学科进行标注和统计,结果见图 3。整体来看,跨学科研究多于单学科研究,占比达 58%;跨学科研究中,大多涉及 2－3 个学科。

提取跨学科研究构建学科交叉网络,见图 4。图 4 中节点大小反映度的大小:节点越大,代表该学科所交叉的学科数量越多。图中共涉及 23 个一级学科,有 96 对学科交叉关系;每个一级学科平均关联 8 个其他学科,且分布较为均衡,整体来看学科间联系较为紧密。

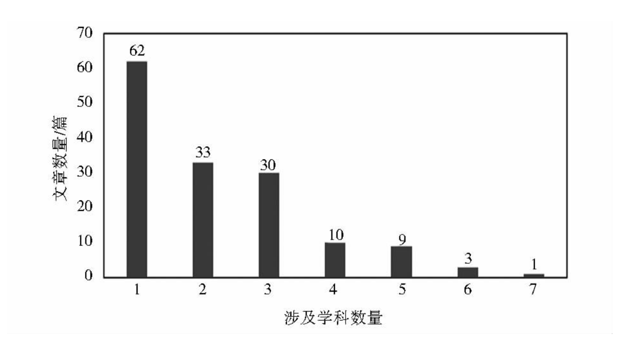


图 3　文章涉及的学科数量分布

表 2 统计了节点的交叉学科数量和平均关联强度。

图书馆、情报与文献学,信息科学与系统科学和计算机科学技术作为数字人文中“数字”学科的代表,位于网络中心,呈现“三足鼎立”之势。这些学科不仅关联学科数量较多,与其关联学科的联系也较为紧密,往往扮演“技术支撑”角色,但侧重不同。图书馆、情报与文献学关联学科数量最多,作为数据资源和数字工具的主要提供者,在多个学科中渗透应用。而信息科学与系统科学学科平均学科关联强度最强,尤其是与图情档学科和计算机学科的联系最为紧密。历史学、文学和社会学作为数字人文中“人文”的主要代表,关联学科较多,但大多关联强度不高;三者之间彼此交叉明显,产生了不少优秀研究[19－20]。图中还显示了若干个分布在网络外围的学科,可以视为数字人文的拓展场景,充分展现了数字人文在广大学科中的发展空间。其中,管理学与经济学虽然关联学科数量不多,但关联强度较强。

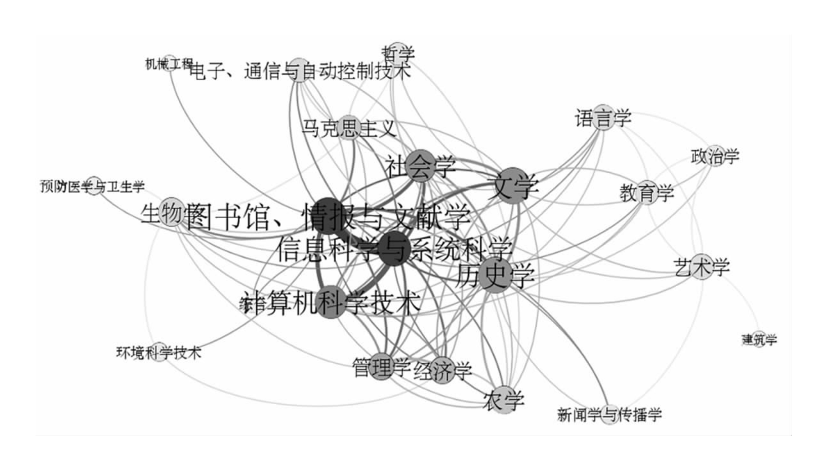


图 4 相关研究学科交叉网络

表 2 各个学科的交叉学科数量和平均关联强度

一级学科	关联学科数量/个	平均学科关联强度
图书馆、情报与文献学	17	6.59
文学	16	3.44
历史学	16	3.25
信息科学与系统科学	15	7.60
计算机科学技术	14	4.50
社会学	13	4.23
农学	10	2.40
生物学	10	1.40
管理学	9	4.56
经济学	9	3.56
马克思主义	8	2.75
语言学	8	1.38
艺术学	8	1.25
电子、通信与自动控制技术	7	1.71
教育学	7	1.29
哲学	6	1.33
政治学	5	1.00
新闻学与传播学	4	1.25
环境科学技术	3	1.00
综合	3	1.00
预防医学与卫生学	2	1.00
机械工程	1	1.00
建筑学	1	1.00

注：平均学科关联强度 = 某一学科与其他学科的共现总次数/该学科的共现学科总数，反映该学科与其关联学科的平均关联强度

3 “数字”运用情况分析与讨论

3.1 数据资源的使用

数字人文的数据是各种数字文件，包括各类人文研究对象的数字化及其元数据。文献分析显示，相关研究主要涉及三类数据来源：专题数据库、自整理数据和综合数据源。

3.1.1 专题数据库

专题数据库，是指针对某一特定主题建立的数据库，此处的数据库是广义概念，包含数据库、数据集或者数据平台，一般具有围绕特定专题、数据规模大、数据质量高、开放可获取等特征。按照建设牵头或主导机构的性质，主要包括国家政府部门主导建设的专题数据库、高校等研究机构主导建设的专题数据库、企业自建专题数据库等类型。从选题视角上看，大多数据库有相对明确的、颇具特色的学科主题，通常与主办机构的性质或研究范畴高度相关，主要用以支持研究团队的长期研究。其中，既有专题范围相对较小的，例如：敦煌研究院数字中心的敦煌学数据库、暨南大学的华人华侨专题数据库、海南大学图书馆的更路数据库[21]、东北大学的张学良史料数据库[22]等。也有专题相对宽泛的，如使用最为广泛的中国近现代思想史专业数据库、中国历代人物传记资料库（CBDB）、名人手稿档案库等。

专题数据库由于知识聚集度高、可获得性强、数据质量佳、数据规模大，深受研究者青睐，有 57% 的研究均有涉及，不乏较有深度的优秀研究，比如《走向新世界：数字人文视野下中国近代“世界”概念的形成与演变》一文，学者从中国近现代思想史专业数据库（1830 –1930）获取中国近代重要政治思想文献，在文本处理与挖掘的基础上，获得了质量高、规模大的文献数据，对 1899 年“世界”概念涌现后至今的演变过程进行深入分析[23]。本文对利用率较高的专题数据库进行总结，见表 3。

3.1.2 自整理数据

自整理数据，是指研究者围绕特定的研究问题，自行收集和整理的数据。一般来说数据规模小，个人研究色彩较重，大部分服务于相对有限的研究问题。例

表 3　使用频率较高的专题数据库

专题数据库	出现次数	数据库简介
中国历代人物传记库[24]	11	中国历代人物传记资料库(CBDB)是由哈佛大学费正清中国研究中心、台湾"中央研究院"历史语言研究所、北京大学中国古代研究中心及中文在线四方共同合作打造的记录中国古典数据的数字人文资源平台,其整合了中国历史上所有重要的传记数据
上海图书馆人名规范库[25]	3	"人名规范库"目前(2021.02.03)网页显示已搜集 1 338 407 条人物数据,时间跨度从先秦到当代,以上图元数据、上图近现代人名规范档、国图规范档、网络百科、人名大辞典、CBDB 等为基础,呈现人物的人名、籍贯、著作、人物小传等信息
汉典古籍语料库[26]	2	汉典古籍是免费的古籍文库,截至调研时间(2021.02.03)收藏古文 38 529 篇
上海图书馆全国报刊索引数据库[27]	2	《全国报刊索引数据库》系国家文化部立项,上海图书馆承建的重大科技项目。数据库格式严格按照国家有关标准,其著录字段包括顺序号、分类号、题名、著者、著者单位、报刊名、年卷期、所在页码、主题词、摘要等十余项。目前该数据库数据已回溯至 1833 年,年更新量在 50 万条左右
中国近现代思想史专业数据库	2	香港中文大学中国文化研究所当代中国文化研究中心研究开发;现由政治大学"中国近现代思想及文学史专业数据库(1830－1930)"计划办公室提供检索服务

如:为了对德意志群像做描绘与分析,学者对《德意志人物志》的文本进行整理和挖掘[28];为探究春秋时期各诸侯国的战争合作与对抗关系,学者对《左传》战争文本进行信息抽取和标注,得到每次战争的进攻方和防守方[29]等。从某种意义上来说,自整理数据的获取方式是传统人文研究中数据获取方式的延展。在信息技术和数字工具出现之前,学者也在做着"从原始的人文史料中提取信息"的工作,只是大多数是以人工方式处理,所能转化、抽取和分析的数据规模较小、管理与组织较复杂。数字工具的出现将人文学者从这部分繁重的工作中解脱出来,从这个角度来看,无论数字人文是否带来人文学科研究范式的颠覆,都有利于相关研究的开展。

3.1.3　综合数据源

综合数据源,是指那些并非专门为数字人文建立的、开放可获取的数据资源,具有数据规模巨大、数据质量较高、内容主题广泛的特点。可以分为两大类:一是各类文献数据库,如 CNKI、CSSCI、Web of Science 等;二是网络开放资源,如维基百科、百度百科、各种网络社区(如豆瓣网、Friend of a Friend 等),在相关研究中有时以辅助数据源的形式出现。如《觅江南佳馔:多源古今文本数据融合的沪上饮食图谱构建》一文中,作者将爱如生专题数据库中的方志数据和"下厨房"网站中的现代菜谱数据相结合,构建了上海饮食的知识图谱用以分析各个实体间的关联演化[12]。

整体来看,三类数据源相互补充,共同构成数字人文研究的数据资源基础。专题数据库数据规模适中、数据质量较好,且具有一定的针对性,应用最为广泛。自整理数据规模最小,但针对性也最强,服务于特定学者的单点研究,但一般不具开放性。综合数据源数据量大、主题广泛、更新及时、获取容易,但针对性较差,有时需要结合具体问题和其它来源的数据一起使用。

3.2　主要研究类型

关键词是文章作者抽取的、用来概括文章主题和研究内容的词或词组。本部分提取每篇文章中的关键词,共计得到 397 个。针对同义词表述不同的问题,人工对其做规范化处理:例如"主题模型"与"主题建模"表义相似,统一规范化为"主题模型";"文献计量学"与"文献计量"统一为"文献计量";"GIS"与"地理信息系统"统一为"地理信息系统"。处理之后得到关键词 378 个,将出现频次大于 5 的关键词定义为高频关键词,删去"数字人文"这一与主题相关度过高的词,将其余的 71 个高频关键词作为共词分析的基础,如表 4 所示:

表 4　高频关键词列表

序号	关键词	词频	序号	关键词	词频
1	关联数据	48	37	莎士比亚	8
2	可视化	44	38	哲学研究	8
3	知识图谱	37	39	深度学习	8
4	社会网络分析	24	40	清代女性	8
5	马克思著作	24	41	观念史	8
6	CSSCI	24	42	地方历史文献	7
7	元数据	23	43	计算叙事	7
8	知识组织	23	44	文学制图	7
9	地理信息系统	19	45	主题	7
10	特色资源	19	46	唐长安	7
11	知识发现	17	47	唐传奇	7
12	国际图像互操作框架(IIIF)	16	48	知识生产	7
13	语义	13	49	研究热点	7
14	知识库	13	50	图书馆	7
15	古籍数字化	12	51	地理叙事	7
16	学术影响力	12	52	空间叙事	7
17	空间人文	12	53	知识关联	7
18	知识模型	12	54	话语	7
19	文献计量	12	55	女性人物	7
20	方志物产	12	56	停用词	6

（续表 4）

序号	关键词	词频	序号	关键词	词频
21	文本挖掘	12	57	明代进士	6
22	民国时期文献	11	58	智能知识服务	6
23	本体	11	59	唐诗知识图谱	6
24	数据库	11	60	人文计算	6
25	古文信息处理	10	61	条件随机场	6
26	知识服务	10	62	文献辨析	6
27	数据可视化	10	63	河南省	6
28	命名实体识别	10	64	人工智能	6
29	CiteSpace	10	65	人文学者	6
30	文本可视化	9	66	人文数据库	6
31	主题模型	9	67	文本关系	6
32	顾太清	8	68	董其昌	6
33	茶文化	8	69	技术责任	6
34	秋红吟社	8	70	网络文档	6
35	在华外文报纸	8	71	自然语言处理	6
36	先秦典籍	8			

由关键词共现频次生成共词矩阵，在此基础上生成相异矩阵，计算方法如下：

$$\text{Dissimiliary}_{a,b} = 1 - \frac{cooccur_{a,b}}{\sqrt{occur_a} \times \sqrt{occur_b}}$$

其中，$\text{Dissimiliary}_{a,b}$表示 a、b 两个关键词的相异系数，$cooccur_{a,b}$ 表示 a、b 两个关键词的共现次数，$occur_a$ 表示 a 词出现的频次，$occur_b$ 表示 b 词出现的频次。使用 SPSS 进行二维尺度分析（ALSCAL），选取平面对称的图形描述关键词的研究结构，对关键词相异矩阵进行二维尺度分析得到相关可视化结果，如图 5 所示。图中横轴代表“数字”维度，数值越大代表研究中的“数字”色彩越重，纵轴代表“人文”维度，数值越大代表研究中的“人文”因素越多。从结果图来看，关键词分布在三个区域，代表了不同类型研究的知识结构，表明不同流派、领域或学术共同体在相关研究中的位置与关系，可用于分析各类研究的研究路径[30]。

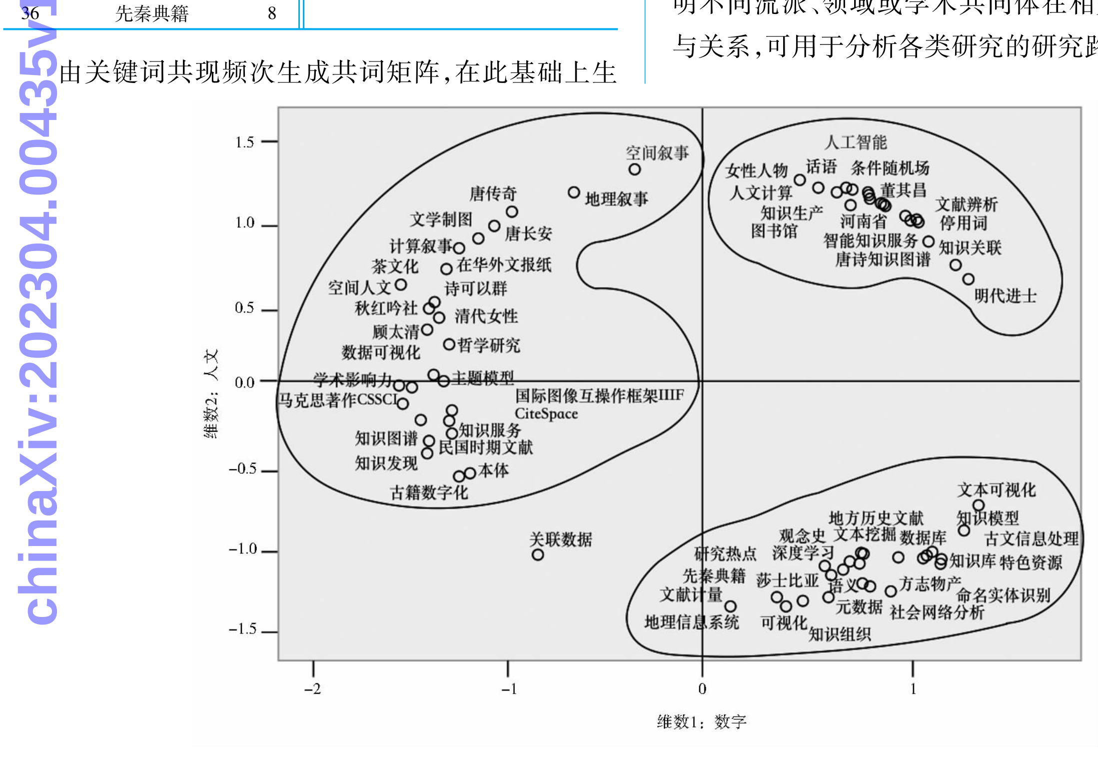


图 5　多维尺度分析结果

3.2.1　资源建设型研究

资源建设型研究是以数据基础设施建设为主的研究，致力于研究各种数字文件的生产与组织，构建独立于具体应用和研究领域，能够长期保存、共建共享和开放利用的数据资源，包括数据库、数据集、数据平台等各种形式。从图中来看，此类研究共包含 23 个关键词，主要分布在右下角。部分关键词与数据组织管理有关，如知识模型、知识组织、数据库等，涉及的具体研究如：以特色馆藏为基础，构建民国文书、方志古籍、地方非遗等特色数据库；梳理领域知识和词表，构建特定领域的元数据、本体和叙词表等；利用馆际联盟开展数据资源融合，为不同专题数据库的关联融合提供数据基础。从分布位置来看，它们在“人文”横轴的下半部分，“数字”横轴的右半部分；人文色彩较少，且从左至右，技术属性越来越强，如文本可视化、知识模型、命名实体识别、古文信息处理等；也有少数研究是学者结合数据资源对特定人文问题的辨析。

资源建设型研究是数字人文领域中起源最早、发展最为成熟的一类研究，也是后续研究开展的基础，这与柯平等学者[3]的研究阶段划分一致。此类研究技术

石静，李阳.数据和技术运用视角下我国数字人文"数字型"研究现状及启示[J].图书情报工作,2021,65(21):141－150.

属性较强,研究者多集中在图书馆、情报与文献学学科、GLAM 等机构,人文学者一般仅在领域知识和建设需求中提供辅助。此外,此类研究还具有实践先于理论的特点,早期很多研究都是先开展实践,再总结理论。但随着数据资源的激增,这种方式带来的信息孤岛、重复建设等问题也日渐凸显,越来越多的学者提出理论和标准建构的重要性。

3.2.2 工具驱动型研究

工具驱动型研究是指将相对成熟的工具、模型或方法应用到人文新场景下的研究,比如 Citespace、主题模型等,研究数量较多。此类研究包括 27 个关键词,从位置来看,主要位于"数字"纵轴的左侧,沿着纵轴向下,关键词的技术属性增强、人文属性减弱,如本体、数字化等;且关键词分布横跨"人文"横轴,较为均衡分散。相比资源建设型研究,工具驱动型研究整体的"人文"色彩更强,与人文学科的融合也更加深入。

由于所使用工具技术属性的强弱程度、人文思想的渗透程度不同,此类研究的最终面貌存在一定差别。比如在横轴下半部分的研究,涉及关键词"Citespace""CSSCI""古籍数字化"等,技术上较为简单,并且人文思想渗透程度较浅,一般多是提供问题场景。而横轴上半部分的研究,涉及关键词"空间人文""计算叙事""地理叙事"等,人文思想的渗透程度较强,不仅提供工具的使用场景,还为技术工具的使用提供目标指引,使研究工具为研究问题服务。值得注意的是,分布位置偏上的时空人文相关研究占比较大,这部分研究不仅用到数字人文史料,还涉及时间和空间位置信息的抽取,在时空维度上考察特定人文问题,是兼具"数字"与"人文"思维的研究成果。如从唐诗中抽取主要空间方位词,探讨唐代诗人乐园空间认知的变化[19];结合人物传记数据和地理位置数据,探究明代进士群体的时空分布与变迁过程[16];从空间视域下考察社会经济和基础文教对作家地理分布的影响[31]等。

整体来看,工具驱动类研究一般均兼具"数字"与"人文"属性,但技术工具的使用水平存在一定差异,这也在一定程度上区分了研究层次;而工具的使用水平主要受到研究中"人文"思维渗透程度的影响。此外,此类研究的主题受数字人文研究热点的影响较大,比如早期研究中数据需求大多是单来源、文本型的,因此文本挖掘技术是研究重点;之后随着研究的深入,数据需求转为多来源、多类型的,相应的研究重点就转为知识融合;近年来,很多人文学者开始关注图像、古汉语数字资源利用,专门的图像处理与古文处理技术也随之大量涌现。

3.2.3 场景融合型研究

场景融合型研究是由人文学科的具体研究情境为主导,综合运用数字资源和数据思维对特定人文问题进行深度辨析的研究。此类别共涉及 21 个关键词,从分布位置来看,处于坐标系的对角线,是三大类型中将"数字"与"人文"融合得较深入的一类研究。

从内容来看,此类研究使用的技术较前两类更加专深,如知识图谱、条件随机场、人工智能等;人文思维的渗透也更深入,大多并非直接享受数据资源建设的成果,而是涉及到数据基础资源建设的细节处理过程,如文献辨析、停用词处理等。这反映出当前数字人文"数字型"研究中,已经出现"数字"与"人文"深度碰撞融合的研究路径。此类研究中,人文学者的角色不单是数据和和工具的使用者,同时也是数据基础设施的建设者。这种角色的转变会对具体研究过程产生影响,使得人文学者从数据资源、数字技术的获取应用环节前置到建设开发环节。这种全过程的深度参与,是数字人文研究深化发展的时代要求,也是重要的未来趋势。

场景融合型研究的典型特点是整个研究过程以人文问题和场景需求为主导,人文眼光和数据思维始终交织存在。虽然早期研究多是简单的数据工具和方法在人文领域的迁移应用,"数字"大过"人文";但是近年来,随着研究的深入和人文学者信息素养的提升,媒介融合时代的数据技术和符号属性已经在一定程度上改变了人文学科故事组织的逻辑,越来越多的人文学者开始注重基于数字资源或利用数据思维去探究人文问题;信息管理等其它学科的学者也不再仅仅"为建库而建库",而是更多地结合人文思维与问题场景开展数据建设。

4 对数字人文研究的若干思考

4.1 面向数字人文研究的数据基础设施搭建问题

科学研究离不开数据的支持,海量的、规范的、开放的、关联的数据资源同样也是数字人文研究的基础"血液"。尽管目前支持数字人文研究的数据库等已经得到了较好的改进,但仍然有以下亟待解决的问题:

(1)数据孤立,重复建设。虽然目前面向数字人文的专题数据库建设如火如荼,但由于缺乏统一的规范和标准,无法互通互享,应用最多最广的只是少数几个大型专题数据库,仍有大量的数据库处于不可用或未开放状态,一方面会造成数据孤立、资源浪费;另一

方面,数据关联取用困难和数据需求增长之间的矛盾,又会带来同类数据的重复建设问题。

(2)数据来源较单一,以文本数据为主。从研究主题来看,当前已经出现了计量电影、时空人文等研究,学者对非文本数据的需求增加,而目前可用的数据资源大多仍来自文本数据。随着数字人文向其他领域的快速渗透,数据资源类型逐渐丰富也是必然趋势。但那些现存的非结构化、以各种媒体形式存在的数据如何处理仍是一个难点,尚且依赖于未来技术的突破。

(3)数字工具的开发不足。数据使用复杂性往往会降低人文学者的使用积极性[32],多数资源建设的重点仍停留在数据收集与知识组织层面,对知识服务的关注不足;当前相关研究中应用到的工具多是从其他学科直接搬运,比如直接将 Citespace、社会网络分析等迁移到人文场景中,缺乏针对特色数据库、特色研究问题的专门工具,这也是导致数据可利用性低的原因之一。

这些问题的解决有赖于各方协同配合,尤其是以信息管理为主的图情档学科更应当积极发挥作用。如有学者就提出图书馆学是数字人文产生的基础学科,应当承担起探索数字人文理论与实践的工作[3]。实际上,目前很多机构也都在关注并重视数字人文基础设施搭建,但更多采用的是独立开发模式,个别采用的众包等模式有待进一步推广。

4.2 数字人文学者的数字学术能力问题

从人文学科的特征来看,其研究的领域特征和个人色彩较明显,没有通用套路,特定研究问题对数据、工具、平台的需求千差万别。如史学研究中需要利用数字技术,对大规模历史文献中复杂的网络关系进行提取和组织,而哲学研究中可能更关注概念的产生与演化。因此,在数字人文相关实践过程中,从数据的获取、选择、分析到呈现的全部环节,都需要特定场景下的人文学者全程紧密参与指导,自然也对其数字学术能力提出了更高的要求。

从研究现状与趋势来看也是如此,资源建设型和工具驱动型研究占比较高,而这两者都与数字技术紧密关联,单纯利用数据资源建设成果和较为简单的数字技术工具已不足够;从未来趋势看,“数字”对于“人文”的作用绝不仅是简单的工具应用,而是超越了数据资源、工具甚至方法[4]。场景融合型研究更是对人文学者的信息素养提出了更高的要求,不仅需要人文学者参与到数据资源建设、技术工具开发过程,还需要其数据思维乃至数据素质的全面提升,能够根据问题情境和研究过程灵活运用数据和工具,达到“数字”与“人文”深度融合。

4.3 数字人文中“人文”的角色转变问题

如前所述,数字人文的热潮引起了学者们的冷思考,关于“数字”与“人文”孰轻孰重的探讨和分歧一直存在,数字人文中人文的定位如何? 数字人文对于人文学科的发展有何益弊? 诸如此类问题层出不穷。

数字人文无法取代传统人文,几乎已经成为共识。数字人文是人文学科的延伸,而非替代,它为人文学者提供方便利用的数据资源和数字工具,一定程度上将其从繁重的资料查找组织工作中解放出来。更重要的是,它为人文学者提供了更广阔的思辨空间,随着数字人文发展过程的深入,人文学者的数字素质得到提升,就能够更好地利用数字反哺传统人文,真正实现“数字”与“人文”的双向共生。当然,数字人文的火热与可观的成果,也无法避免会在一定程度上减弱人文学者对其它方面问题的关注,对人文科学的思辨性和深度产生负面影响,造成“重数字,轻人文”的现象。归根结底,数字人文仍是人文色彩为主的领域,“数字”一词具有明显的工具属性(当然并不仅仅是工具),工具本身无好坏,其产生的效果如何依赖于使用者如何运用。倘若使用者自身存有“唯数据论”的思维误区或数据能力欠缺的问题,最后结果便很可能是弊大于利。但因噎废食之举不可取,降低消极影响的方法应当是提高使用者的能力,而非抛弃工具。此外,人文学科的“数字”转化也是未来趋势,因为从根本上来说,数字人文的出现并不是计算机、信息等强势学科对人文学科的辐射,而是来自于人文学科本身研究基础设施和学术环境的变化[34]。

从资源建设、工具驱动到场景融合,“数字”与“人文”的碰撞正在不断加深。在这个过程中,也必将带来各类研究路径的演化,人文学者需要在各类型研究中适当增加参与度,这不仅有利于这两类研究的发展,也能通过数据资源质量改进、技术工具适用改善、学者数字学术能力提升等方式,间接将数字人文融合研究带上新的台阶。因此,人文学者既应该保持批判与警醒,更应该拥抱新的机遇与挑战,在人文精神和数据思维中寻求平衡。

5 结语

本文以数字人文中的“数字型”研究为研究对象,探究数字人文研究中“数字”与“人文”交叉融合的问题,从研究概况到“数字”运用情况、主要研究类型,再到深度反思,得到了一些结论和启示:

(1)图情档学科、文史学科和计算机学科是主要的研究阵地，且逐渐向社会学、经济学等领域渗透，凸显出较强的交叉性；整体没有明显的“一家独大”现象，而是分布均衡、百家争鸣。各个学科、各方机构在相关研究中都有广阔空间，因此利用自身特色与优势，在数字人文领域打造自身学科话语，目前处于“窗口期”。

(2)少数几个大型专题数据库和研究者自整理数据是主要数据来源，大多专题数据库的利用率不高。一方面是由于数据本身质量不佳、可获取性低，另一方面也受限于研究者的数据素养。未来可尝试通过机构联合或学科联合，自上而下对各方“闲置”资源再梳理、再整合、再利用，同时在此过程中增强用户意识，关注研究者的意见，使得专题数据库真正服务于学术研究。

(3)相关研究主要有三类：资源建设型、工具驱动型和场景融合型，其中场景融合型研究兴起相对较晚、数量较少，但其在“数字”和“人文”中获得了较好的平衡和深度的融合；从二元尺度分析结果看，其与工具驱动型研究的边界逐渐靠近，随着研究的积累和深入，场景融合型研究有望成为工具型研究的未来趋势，并且影响资源建设型研究的热点走向。从研究者来看，则需要转换思维，以场景意识为主导，巧妙挖掘数据，适当利用工具，才能在新趋势下抓住机遇，提升数字人文相关研究品质。

当然，本文也有一定的局限：①仅以文献为研究的数据来源，可能对相关研究覆盖不全，尤其是一些社会机构建立的数字人文专题数据库及行业报告并未涵盖在内。②在文献的选择和编码标注过程中，也可能存在一定的主观性，这可能对结果会产生一些偏差。因此，未来研究中可考虑拓展为基于多源数据的分析；同时，也可进一步增加对国外文献的综述，对国内外数字人文发展情况进行对比分析，为数字人文及其相关学科发展提供参考。

参考文献：

[1] KIRSCHENBAUM M G. What is digital humanities and what's it doing in English departments?[J]. ADE bulletin, 2010(120): 55－61.
[2] 李振奚，建武. 信息化进程的前沿探索——“数字化与21世纪人文精神”研讨会综述[J]. 探索与争鸣, 2003(2):34－36.
[3] 柯平，宫平. 数字人文研究演化路径与热点领域分析[J]. 中国图书馆学报, 2016,42(6):13－30.
[4] 王丽华，刘炜，刘圣婴. 数字人文的理论化趋势前瞻[J]. 中国图书馆学报, 2020,46(3):17－23.
[5] 黄敏聪. 基于人文计算的汉字简繁体演变定量分析[J]. 科技视界, 2012(18):64－67.
[6] 曾子明，周知. 面向数字人文的图像语义描述模型研究[J]. 情报理论与实践, 2018,41(1):116－121.
[7] 宋宁远，王晓光. 基于情节本体的叙事性文本语义结构化表示方法研究[J]. 中国图书馆学报, 2020,46(2):96－113.
[8] 周莉娜，洪亮，高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作, 2019,63(2):24－33.
[9] 童茵. 语义化知识模型构建与关联数据研究——董其昌数字人文数据绘制报告[J]. 科技传播, 2020,12(5):135－137.
[10] 杨佳颖，邓璐芗，许鑫. 觅江南佳馔：多源古今文本数据融合的沪上饮食图谱构建[J]. 图书馆论坛, 2020,40(10):24－31.
[11] 王晓光，侯西龙，程航航，等. 敦煌壁画叙词表构建与关联数据发布[J]. 中国图书馆学报, 2020,46(4):69－84.
[12] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报, 2016,42(2):66－80.
[13] 杨海慈，王军. 宋代学术师承知识图谱的构建与可视化[J]. 数据分析与知识发现, 2019,3(6):109－116.
[14] 邓君，孙绍丹，王阮，等. 数字人文视阈下明代科举进士群体时空网络结构分析[J]. 图书情报工作, 2020,64(17):4－17.
[15] 郭林斐，刘广钟. 基于Neo4j不确定性数据处理技术的研究[J]. 计算机技术与发展, 2020,30(1):25－31.
[16] 梁继文，江川，王东波. 基于多特征融合的先秦典籍汉英句子对齐研究[J]. 数据分析与知识发现, 2020,4(9):123－132.
[17] 邱伟云，严程. 心寄乐园，凝望人间：中唐诗空间方位的数字人文研究[J]. 西南民族大学学报(人文社科版), 2020,41(8):161－171.
[18] 邱伟云，金观涛，刘青峰，等. 中国近代平等观念形成之数字人文研究：以报刊为中心[J]. 清华大学学报(哲学社会科学版), 2018,33(6):1－33.
[19] 李文化，陈虹，李冬蕊. 数字人文视域下的南海更路簿综合研究[J]. 大学图书馆学报, 2020,38(2):91－98.
[20] 姚天泓，陈艳梅，刘革，等. 基于CIDOC-CRM的数字人文史料资源语义化知识组织研究——以张学良史料资源为例[J]. 图书馆学刊, 2019,41(7):35－43.
[21] 邱伟云，郑文惠. 走向新世界：数字人文视野下中国近代“世界”概念的形成与演变[J]. 南京大学学报(哲学·人文科学·社会科学), 2020,57(5):88－106.
[22] UNIVERSITY H, SINICA A, UNIVERSITY P. China Biographical Database (CBDB)[EB/OL]. [2021－02－03]. https://projects.iq.harvard.edu/cbdb.
[23] 上海图书馆人名规范库[EB/OL]. [2021－02－03]. http://names.library.sh.cn/mrgf/service/person/list.
[24] 汉典古籍语料库[EB/OL]. [2021－02－03]. http://gj.zdic.net/.
[25] 全国报刊索引数据库[EB/OL]. [2021－02－03]. https://www.cnbksy.com/search/advance.
[26] 王涛. 数字人文框架下《德意志人物志》的群像描绘与类型分

析[J]. 历史研究, 2018(5):148-166.
[27] 范文洁, 李忠凯, 黄水清. 基于社会网络分析的《左传》战争计量及可视化研究[J]. 图书情报工作, 2020,64(6):90-99.
[28] 张勤, 马费成. 国外知识管理研究范式——以共词分析为方法[J]. 管理科学学报, 2007(6):65-75.
[29] 彭霞, 刘敏, 杨莉, 等. 空间视域下明清女性作家地理分布及其成因——以松江府为例[J]. 农业图书情报学报, 2020,32(9):31-38.
[30] 赵雪芹, 莫长镭, 李天娥, 等. 人文学者的数字人文接受意愿影响因素——以历史学者为中心的考察[J]. 图书馆论坛, 2020,40(4):40-49.
[31] 陈海珠, 包平. 创新人文研究范式的数字人文——基于数字人文在史学研究领域应用渗透的文献考察[J]. 图书馆理论与实践, 2017(3):11-16.

作者贡献说明:
石静:主要负责数据分析与内容撰写;
李阳:主要负责逻辑梳理与内容修改。

Present Situation and Enlightenment of Digital Humanities "Digital-type" Research in China from the Perspective of Data and Technology Application

Shi Jing Li Yang
School of Information Management, Nanjing University, Nanjing 210023

Abstract: [**Purpose/significance**] Many research findings in digital humanities have emerged in China, and digital humanities develops gradually from the cognitive exploration stage to the digital exploration stage in recent years. From the micro perspective of "data resources" and "digital technology", this paper describes the actual panorama of the collision between "digital" and "humanity" by quantitative analysis, in order to have an intuitive and practical cognition of "digital-type" research in digital humanities. [**Method/process**] By using bibliometrics, qualitative analysis and multi-dimensional scale analysis, this paper made an in-depth analysis and reflection on domestic digital humanities related research from the aspects of research institutions distribution, interdisciplinary characteristics, "digital" utilization and primary research types. [**Result/conclusion**] From the perspective of research institutions, digital humanities research is evenly distributed and has broad space in multiple disciplines; from the perspective of data, a few large thematic databases and data self-collated by researchers are the main sources; from the perspective of research types, there are resource building research, tool-driven research and scenario-integrated research, each of them has a specific focus, and scenario-integrated research achieves a better balance between "digital" and "humanity". In the future, it is necessary to actively build appropriate data infrastructures, cultivate the digital academic ability of scholars, and promote the transformation of the role of "humanity" to support the paradigm change of digital humanities research.

Keywords: digital humanities interdisciplinary data resources new technique humanity value